Frontier Topics in Empirical Economics: Week 13
Peer Effect and Spillover

Zibin Huang [1]

[1] College of Business, Shanghai University of Finance and Economics

December 15, 2024

Introduction

- In this week, we are going to investigate an important empirical question
- How to identify and estimate peer effect/spillover effect?
- People may think it is straightforward and simple
- Just run y on $\bar{y}$ or $\bar{x}$
- But actually it is very complicated and dangerous!

Introduction

- We will discuss this issue from two perspectives
  - Technically: Identification and inference failure
  - Intuitively: Interpretation of the peer effect coefficient
- The related MHE chapter is 4.6.2
- However, it is not so detailed
- I recommend you to read the original paper of Angrist (2014) and Manski (1993)

Introduction

- Let me give you a brief preview of the conclusion
- First, you can never distinguish among endogenous effects, exogenous effects, and correlated effects: Reflection problem
- Second, never run regressions like y on $\bar{y}$ for the same group!
- Third, when running y on $\bar{x}$:
  - Make sure group formation is random or quasi-random
  - Check all possible alternative channels that can drive this results such as measurement errors
- Fourth, separate people affecting others from people being affected
  Group for $\bar{y}$ and $\bar{x}$ is different from group for y

Reflection Problem

- Peer effects are intrinsically very difficult to identify
- Because it is hard to distinguish among behavior causation, characteristics causation, and common environment
- Manski (1993) named this ”reflection problem”

Reflection Problem

- When you see co-movements of a person and his image in a mirror
- Without knowledge of optics, how can you differentiate between:
  - The person's movements cause the movements of the image
  - Some external stimulus causes person and image to move together

Reflection Problem

- In general, individuals in the same group tend to behave similarly for the following three reasons:
    - Endogenous effects: an individual's behavior is affected by the behaviors of the group
    - Exogenous (contextual) effects: an individual's behavior is affected by the exogenous characteristics of the group
    - Correlated effects: individuals in the same group have similar characteristics or face same institutional environments

Reflection Problem

- Let's take students in a classroom as an example
- Why do we see similarity of bullying behavior for students in the same class?
  - Endogenous effects:
    A students bullies others because his/her friends do so
  - Exogenous (contextual) effects:
    A students bullies others because his/her friends come from violent families
  - Correlated effects 1:
    All students in this class bully others because they all come from families with violent fathers
  - Correlated effects 2:
    Students in this class bully others because their head teacher does not care

Reflection Problem

- Endogenous/Exogenous effects are different types of spillovers
- Correlated effect is purely a contamination
- Unfortunately, it is generally impossible to identify these three effects separately
- Even in a random/quasi-randomization environment
- Let's see why this is the case

Reflection Problem

- Denote y as a scalar outcome, e.g. a student's test score
- x as group attribute, e.g. class indicator
- z as observed attributes that directly affect y, e.g. family SES
- u as unobserved attributes that directly affect y, e.g. teacher ability
- Consider the following equation:

$$y = \alpha + \beta E(y|x) + E(z|x)'\gamma + z'\eta + u \tag{1}$$

- We assume that $E(u|x, z) = x'\delta$, a CIA quasi-random setting
- Unobserved terms can be absorbed in class FEs

Reflection Problem

- Take the conditional expectation w.r.t. x and z:

$$E(y|x,z) = \alpha + \beta E(y|x) + E(z|x)'\gamma + z'\eta + x'\delta \tag{2}$$

- β is the endogenous effect
- γ is the exogenous effect
- δ is the correlated effect
- Can we identify all of them separately?

Reflection Problem

- Observe that we have conditional expectation of y on both side
- We then take expectation w.r.t. z for both sides:

$$E(y|x) = \alpha + \beta E(y|x) + E(z|x)'\gamma + E(z|x)'\eta + x'\delta \tag{3}$$

- $E(y|x)$ solves this "social equilibrium" equation:

$$E(y|x) = \alpha/(1-\beta) + E(z|x)'[(\gamma+\eta)/(1-\beta)] + x'\delta/(1-\beta) \tag{4}$$

Reflection Problem

- Inserting (4) into (2):

$$E(y|x,z) = \alpha/(1-\beta) + E(z|x)'[(\gamma+\beta\eta)/(1-\beta)] + x'\delta/(1-\beta) + z'\eta \quad (5)$$

- Using a linear regression, we can identify $\alpha/(1-\beta)$, $(\gamma+\beta\eta)/(1-\beta)$, $\delta/(1-\beta)$, and η separately
- But that's it. Nothing more we can do.
- Four reg coefficients, five unknowns
- Can we distinguish between these three effects? No.

Reflection Problem

- Therefore, Manski (1993) proves that in general, we cannot distinguish between endogenous effect, exogenous effect, and correlated effect.
- This is disappointing. Can we still identify some meaningful spillover effect?
- The only hope is that we give up on decomposing everything
- Rather, we identify some simple composite effect
- Ignore the effect of $\bar{y}$ when running y on $\bar{x}$
- Or consider only y on $\bar{y}$, but not $\bar{x}$

Reflection Problem

- However, even in this case, we have to be very careful
- Let's go to Angrist (2014) to see why

The Perils of Peer Effects

- We have shown that distinguish different peer effects carefully is not feasible
- Can we identify either endogenous or exogenous peer effect taking the other as "channel"?
- For instance, we run y only on $\bar{y}$ or $\bar{x}$, rather than both of them
- What is the interpretation of these coefficients?
- Let's analyze them carefully.
- It is not as straightforward as you may think: Angrist (2014)

The Perils of Peer Effects: Reg y on $\bar{x}$

- There are two kinds of peer effect regressions
- We can focus on exogenous effect and regress y on $\bar{x}$
- We can focus on endogenous effect and regress y on $\bar{y}$
- Let's discuss them one by one

The Perils of Peer Effects: Reg y on $\bar{x}$

- First, a good example for exogenous effect is social return of education
- What is the impact of province-level average education on an individual's wage?
- Then we can directly run the following regression:

$$Y_{ij} = \mu + \pi_0 s_i + \pi_1 \bar{S}_j + \nu_{ij} \tag{6}$$

- Y_{ij} is the wage of individual i in province j
- s_i is the education level of individual i
- $\bar{S}_j$ is the average education of people in province j

The Perils of Peer Effects: Reg y on $\bar{x}$

- It can be shown that we can express π_1, π_1 as follows:

$$\pi_0 = \rho_1 + \phi(\rho_0 - \rho_1) \tag{7}$$

$$\pi_1 = \phi(\rho_1 - \rho_0) \tag{8}$$

- ρ_0 is the regression coefficient for a reg of Y_{ij} on s_i
- ρ_1 is the regression coefficient for a 2SLS regression:
  Y_{ij} is the outcome, s_i is the endogenous variable, group dummies $I(j)$ are the instrument
- $\phi = \frac{1}{1-R^2} > 1$ is a dummy related to first stage R^2 for the 2SLS

The Perils of Peer Effects: Reg y on $\bar{x}$

- How to interpret this?

$$\pi_0 = \rho_1 + \phi(\rho_0 - \rho_1) \tag{9}$$
$$\pi_1 = \phi(\rho_1 - \rho_0) \tag{10}$$

- We care about spillover effect π_1
- It is positively related to ϕ and $(\rho_1 - \rho_0)$
- As long as $(\rho_1 - \rho_0) \neq 0$, we will have a non-zero π_1

The Perils of Peer Effects: Reg y on $\bar{x}$

- How to interpret this?
- As long as there is a difference between the estimates of:
  - An OLS reg of Y_{ij} on s_i
  - A 2SLS reg of Y_{ij} on s_i using group dummies $I(j)$ as IV
- You will have a non-zero $\pi_1 \Rightarrow$ non-zero "peer effect/spillover" estimate

The Perils of Peer Effects: Reg y on $\bar{x}$

- How to understand this in our context?
- In the OLS regression, it pins down the correlation between your own wage and your own education
- In the IV regression, consider different provinces are randomly assigned Compulsory Education Laws (CDL)
- Then OLS regression underestimate the effect of education on wage when peer effect is there
- Because an increase in i's education can promote wage for not only i and other people without education increase (control in the same province)

The Perils of Peer Effects: Reg y on $\bar{x}$

- Meanwhile, IV regression essentially compares results from different provinces
- Whose variations are driven by the randomly assigned CDL
- This will not be affected by the peer effect (spillover happens within province)
- Subtracting IV by OLS gives you peer effect

The Perils of Peer Effects: Reg y on $\bar{x}$

- However, is spillover the only reason why IV result is deviate from OLS?
- Of course NOT!
- There can be many reasons why you have a difference between the estimates of OLS and 2SLS regressions!
- Selection bias, measurement error...
- For example, if selection bias exists, OLS can overestimate the results
- Or, a classical measurement error in education leads to attenuation bias in OLS
- This can also create the gap between OLS and IV estimation

The Perils of Peer Effects: Reg y on $\bar{x}$

In general, we have the following implications:

- Peer effect is not essential for the existence of the difference
- It means that even if you detect a non-zero coefficient in regression (6), it can be due to selection bias or measurement error
- Even if real peer effect exists, the results of regression (6) can be contaminated by many other reasons

The Perils of Peer Effects: Reg y on $\bar{x}$

- Do we have any method to alleviate this issue?
- Not so much we can do for the existence of selection bias
- But we can test whether the "peer effect" actually comes from measurement error
- It requires a simulation process used in Carrell, Hoekstra, and Kuka (2018) and Feld and Zölitz (2017)

The Perils of Peer Effects: Reg y on $\bar{x}$

- The basic idea is simple
- We are worried that the detected coefficient π_1 in regression (6) is due to the measurement error
- Then a simple implication is that:
  If we create more measurement error, magnitude of π_1 would be larger
- Assume that we proxy education level by college degree attainment
- Further assume that we have sample size N, with x% of the sample being college educated

The Perils of Peer Effects: Reg y on $\bar{x}$

- We implement the following simulation process to exclude the measurement error contamination
  - (1) Randomly select $p\%$ of the sample
  - (2) In the selected sample, randomly assign $x\%$ individuals to have college education (replace their true education in data)
  - (3) Run the main regression with this fake data
- We repeat this process while varying p from 0% to 100%
- 0% means the baseline estimates without any added measurement error
- 100% means the extreme case when all observations are measured with error

The Perils of Peer Effects: Reg y on $\bar{x}$

- If we find that the estimated coefficient π_1 grows larger and larger when we add in more and more noise
- Then, measurement error may be an important reason for the detected coefficient
- If it is not, then it is likely that π_1 comes from true peer effect but not measurement error

The Perils of Peer Effects: Reg y on $\bar{x}$

An important tip:

- Full randomization to groups like RCT CANNOT solve this issue
- It is not about the randomization of $\bar{S}_j$
- It is about why results of these two regressions can be different
  - An OLS reg of Y_{ij} on s_i
  - A 2SLS reg of Y_{ij} on s_i using group dummies $I(j)$ as IV

The Perils of Peer Effects: Reg y on $\bar{y}$

- We have discussed the case of reg y on $\bar{x}$
- The second case is reg y on $\bar{y}$
- This seems to give us some information about the endogenous effect
- However, this regression is even more dangerous than the first one

The Perils of Peer Effects: Reg y on $\bar{y}$

- To begin with, directly running y on $\bar{y}$ makes no sense
- It will give you a coefficient of 1. Why?
- Consider a school dropout issue
- Let s_{ij} be the dropout decision for student i in school j; $\bar{S}_j$ is the average dropout rate in school j
- We run the following regression:

$$s_{ij} = \mu + \pi_2 \bar{S}_j + \nu_{ij} \tag{11}$$

- The OLS will give you $\bar{\pi}_2 = 1$ for sure

The Perils of Peer Effects: Reg y on $\bar{y}$

- Let me give you a simple proof

$$\bar{\pi}_2 = \frac{\sum_j \sum_i s_{ij}(\bar{S}_j - \bar{S})}{\sum_j \sum_i (\bar{S}_j - \bar{S})^2} = \frac{\sum_j (\bar{S}_j - \bar{S}) \sum_i s_{ij}}{\sum_j n_j (\bar{S}_j - \bar{S})^2} = \frac{\sum_j (\bar{S}_j - \bar{S}) n_j \bar{S}_j}{\sum_j n_j (\bar{S}_j - \bar{S})^2}$$

$$= \frac{\sum_j (\bar{S}_j - \bar{S}) n_j \bar{S}_j}{\sum_j [n_j \bar{S}_j (\bar{S}_j - \bar{S}) - n_j \bar{S} (\bar{S}_j - \bar{S})]} = 1$$

- Note that we have $\sum_j n_j \bar{S}(\bar{S}_j - \bar{S}) = 0$

The Perils of Peer Effects: Reg y on $\bar{y}$

- To avoid the issue we just mentioned, we can run a leave-one-out regression:

$$s_{ij} = \mu + \pi_3 \bar{S}_{-ij} + \mu_{ij} \tag{12}$$

- $\bar{S}_{-ij}$ is the average school dropout rate excluding student i
- The coefficient of this regression is no longer guaranteed to be 1
- However, it is still almost impossible to say we identify some peer effect/spillover

The Perils of Peer Effects: Reg y on $\bar{y}$

- Because any school level random shock can create spurious peer effects!
- For example, a good principal can lead all students in a school not to dropout
- It has nothing to do with peer effect or spillover
- Again we go back to Manski (1993)
- It is almost impossible to distinguish between real peer effects and contamination of correlated effects

The Perils of Peer Effects: Randomization and First Stage

- Except for the issues we just mentioned
- We also need to be very careful about the traditional selection problem
- Usually, grouping is not random
- Good students select to good schools; good employees select to good firms
- Thus, the prerequisite is to have a random/quasi-random group forming, before you start to consider the previous issues

The Perils of Peer Effects: Randomization and Variation of $\bar{S}$

- However, once you have a random group formation, variations of the independent variable can be a problem
- If students are randomly assigned to schools
- For all schools, $E[s_{ij}]$ is the same
- If the number of student is very large in each school, then $\bar{S}_j$ will also be very similar
- But you need variations in $\bar{S}_j$ to identify the peer effect!

The Perils of Peer Effects: Randomization and Variation of $\bar{S}$

- Thus, here you have a tradeoff
- If the grouping is totally random, you may have very small variation in independent variable $\bar{S}$
- If the grouping is not that random, you may have enough variations in $\bar{S}$
- But the selection issue can be severe
- Therefore, in practice, the best case should be:
  - You have a random grouping
  - Meanwhile, the group size is not that large, which gives you enough small sample variation in independent variable $\bar{S}$

The Perils of Peer Effects: Empirical Suggestions

- In general, peer effects are difficult to identify
- Here are some empirical suggestions
- 1. Clearly separate between *subjects* who receive the peer effects and the peers who provide the effect
  - What is the impact of fellow boys' test score on a boy? ×
  - What is the impact of boys' test score on a girl? $\surd$

The Perils of Peer Effects: Empirical Suggestions

- 2. Make sure the fundamental OLS and 2SLS can give you same result in the absence of peer effects
  - We cannot do too much on this
  - One thing you should do is to check the measurement error issue using methods in Carrell, Hoekstra, and Kuka (2018) and Feld and Zölitz (2017)

The Perils of Peer Effects: Empirical Suggestions

- 3. Check the tradeoff between randomization and variation
  - I would always put randomization to be the most important thing
  - Thus, balance check is essential as the first step in peer effect analysis
  - You should run $\bar{S}$ on potential confounders to see whether grouping is random
  - Also, you should check you still have enough variation in $\bar{S}$ after randomization

Application

- The application paper for homework this week is Huang and Zhang (2023)
- This paper investigate the impact of migrant children's school enrollment restriction on education outcomes in China
- There are two parts:
  - Peer effect estimation of migrant/left-behind children on their classmates
  - Spatial equilibrium model to show the overall cost of this discrimination
- This paper helps you to understand how to apply the things we learned in the last two weeks: peer effect + DCM

References

Angrist, Joshua D. 2014. "The Perils of Peer Effects." *Labour Economics* 30:98–108.

Carrell, Scott E, Mark Hoekstra, and Elira Kuka. 2018. "The Long-run Effects of Disruptive Peers." *American Economic Review* 108 (11):3377–3415.

Feld, Jan and Ulf Zölitz. 2017. "Understanding Peer Effects: On the Nature, Estimation, and Channels of Peer Effects." *Journal of Labor Economics* 35 (2):387–428.

Huang, Zibin and Junsen Zhang. 2023. "School Restrictions, Migration, and Peer Effects: a Spatial Equilibrium Analysis of Children's Human Capital in China." *Unpublished Manuscript* .

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies* 60 (3):531–542.